Final Report • 21 December 1992

RESEARCH IN TEXT PROCESSING

SRI Project No. ECU 1592
ARPA Order No. 7331
Contract No.: N00014-90-C-0220

Prepared for:

Lt. Cmdr. Robert Powell
Code 1133, Scientific Officer
Computer Science Division
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217-5000

Prepared by:

Jerry R. Hobbs, Ph.D.
Artificial Intelligence Center
Computing and Engineering Sciences Division

Approved by:

C. Raymond Perrault, Director
Artificial Intelligence Center

Donald L. Nielson, Senior Vice President
Computing and Engineering Sciences Division

Executive Summary

SRI's DARPA-sponsored project on text processing has consisted of a broad range of efforts, from near-term practical system implementation to advanced theoretical research. The research in this project was carried out between September 1990 and November 1992.

We have distinguished two distinct text processing tasks: *information extraction* and *text understanding.* In information extraction,

- Only a fraction of the text is relevant; in the case of the MUC-4 terrorist reports, probably only about 10% of the text is relevant.
- Information is mapped into a predefined, relatively simple, rigid target representation; this condition holds whenever entry of information into a database is the task.
- The subtle nuances of meaning and the writer's goals in writing the text are of no interest.

This contrasts with text understanding, where

- The aim is to make sense of the entire text.
- The target representation must accommodate the full complexities of language.
- One wants to recognize the nuances of meaning and the writer's goals.

The recognition of the distinction between these two tasks was itself a significant result of the project, since it opened the way for practical solutions to the latter task in the very short term.

In accordance with our broad goals, we have focused on three specific areas of research:

- We have pushed the frontiers of the theory of discourse interpretation, especially in the areas of recognizing discourse structure, interpreting metaphors, and recognizing the speaker's plan. This has been done in an integrated framework for text understanding using abduction as the means of drawing inferences from a knowledge base of commonsense background knowledge.

- We have improved upon the TACITUS system for text understanding to bring it to the point where it could handle very complex news reports. The principal area of improvement was in robustness. The improvements include such capabilities as statistical text filtering, handling unknown words, parsing long sentences, recovering from failed parses, and doing abductive inference efficiently.

- We have discovered, implemented, and successfully demonstrated a new method for extracting information from texts for entry into databases, based on cascaded finite-state automata. The system implementing these ideas, FASTUS, is as effective as the best information extraction systems in existence and an order of magnitude faster.

Our achievements in these three areas are described more fully in the next three sections. As we discuss our efforts and achievements in each of these areas, we will cite the relevant papers. The papers are included with, and constitute a part of, this final report.

Much of the work was organized around the MUC-3 and MUC-4 evaluations of message understanding systems, in May 1991 and May 1992, respectively (Sundheim, 1991, 1992). The methodology chosen for these evaluations was to score a system's ability to fill in slots in templates summarizing the content of newspaper articles on Latin American terrorism. The articles ranged from one third of a page to two pages in length. The template-filling task required identifying, among other things, the perpetrators and victims of each terrorist act described in an article, the occupations of the victims, the type of physical entity attacked or destroyed, the date, the location, and the effect on the targets. Many articles described multiple incidents, while other texts were completely irrelevant.

The principal measures in the evaluations were recall and precision. *Recall* is the number of answers the system got right divided by the number of possible right answers. It measures how comprehensive the system is in its extraction of relevant information. *Precision* is the number of answers the system got right divided by the number of answers the system gave. It measures the system's accuracy. In addition, in MUC-4, a combined measure, called the F-score, was used. It is a kind of weighted average of recall and precision.

In MUC-3, SRI used the TACITUS system, as described in Enclosure 7. We scored the highest among 15 sites in precision, although we were somewhere in the middle in recall. In MUC-4, SRI used the FASTUS system.

Only one of 16 sites scored significantly higher than SRI, and FASTUS was an order of magnitude faster than any other comparable system.

1 Theoretical Research in Text Understanding

1.1 Interpretation as Abduction

In SRI's original DARPA-sponsored text understanding project, from 1985 to 1990, we achieved a breakthrough in knowledge-based text understanding with the discovery of the "Interpretation as Abduction" framework. It has long been understood that most of the information conveyed by a text is implicit and must be recovered by a context-sensitive process of inference from a knowledge base of background world knowledge. Over the years, a number of schemes have been proposed for carrying out these inference processes, but they have in general been highly ad hoc and unmotivated. The abductive approach made all of this work fall together in an elegant and compelling integrated framework, in which all knowledge was expressed in a uniform fashion and all inferences were drawn by a single process. This approach is described in "Interpretation as Abduction" (Enclosure 1), which is a significantly expanded version of a paper that was included in our final report in 1990.

The fundamental insight is that the interpretation of a text is the best explanation for the situation it describes. This idea is cashed out procedurally in terms of theorem-proving technology by saying that the interpretation of a text is the minimal proof of the logical forms of the sentences in the text. This much is deduction. But we will not have all the facts we need in the knowledge base to prove the logical form, so we will have to make a minimal number of assumptions. Deductions plus assumptions is abduction.

It turns out that all of the "local pragmatics" problems that have been of concern in natural language processing research simply fall out of this formulation of what an interpretation is. These problems include the resolution of reference and syntactic and lexical ambiguity, the discovery of the specific intended meanings for vague predicates and compound nominals, and the expansion of metonymies and ellipses. In addition, the view of text understanding as schema recognition, which was a common but clearly inadequate previous account, is subsumed under the "Interpretation as Abduction" framework.

The "Interpretation as Abduction" framework can furthermore be combined with the older "parsing as deduction" idea to yield an integrated,

uniform approach to syntax, semantic analysis, and local pragmatics, an approach moreover that gives us a natural treatment of ungrammatical utterances.

This much of the formulation of the theory had been accomplished by the beginning of the current project.

In the current project, we have expanded the theory to encompass three more areas of text understanding—the recognition of discourse structure, the interpretation of metaphors, and the plan ascription problem. In addition, we have made progress on making the process of abduction more efficient and on devising a principled semantics for the weights in the scheme of weighted abduction we use.

1.2 Discourse Structure

The basic idea behind our approach to recognizing discourse structure is that there are a small number of "coherence relations" expressing essentially causal, figure-ground, and similarity relationships. These coherence relations are the relations among the situations described by adjacent sentences or larger segments of discourse, that are conveyed by the mere adjacency of those segments. The interpretation is the best explanation of the information conveyed by the text, including the information conveyed by that adjacency of segments. The best explanation of the adjacency is generally provided by one of the coherence relations. Since the discourse segments are defined recursively, this approach to coherence relations yields a tree-like structure for the entire discourse. This approach to discourse structure is developed more fully in Section 6.3 of "Interpretation as Abduction" (Enclosure 1). An example involving discourse structure and metaphor interpretation is given in Section 4.3 of "Metaphor and Abduction" (Enclosure 2).

1.3 Metaphor

The fundamental process underlying metaphor is mapping between a source domain and a target domain. Basic concepts in the target domain are mapped into basic concepts in the source domain. Inferences are drawn in the source domain to derive complex concepts there. These are then mapped back into complex concepts in the target domain. Any approach to metaphor must spell out how the mapping is accomplished. To incorporate the interpretation of metaphors into the abductive framework we took the

mapping to be effected by ordinary axioms expressing identity between entities in the source and target domains. This work is described in "Metaphor and Abduction" (Enclosure 2). In this paper, three examples are worked out—a conventional spatial metaphor, a category metaphor whose interpretation depends on the discourse context, and a novel metaphor involving mapping between two large-scale schemas. Finally, there is a discussion of the problems that arise from dealing with metaphor in a framework where consistency-checking is one of the principal operations.

1.4 Plan Ascription

The third problem that the abductive approach was extended to handle is the problem of plan ascription. A leading view of utterance interpretation for the last decade and a half has been that the utterance is an action in the speaker's plan for achieving some goal and the hearer must discover the relation between the utterance and the speaker's plan; call this the Intentional Perspective. This view may seem to be in conflict with the "Interpretation as Abduction" view that the interpretation of a text is the best explanation of why it would be true, which may be called the Informational Perspective. The paper "A Unified Abductive Treatment of the Intentional and Informational Aspects of Discourse Interpretation: A Preliminary Report" (Enclosure 3) shows how the plan ascription problem can be handled in the abductive framework and how the Intentional Perspective subsumes the Informational Perspective. This paper will appear as a technical report; the material in it has been included in many of the presentations listed below.

During the period of the previous DARPA contract, Douglas Appelt and Martha Pollack developed another abductive approach to the plan ascription problem. Appelt gave a presentation on this, entitled "Weighted Abduction for Plan Ascription", at the Seminar on User Modeling, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Saarbruecken, Germany, November 1990.

Douglas Appelt has more recently written a critical evaluation of abductive approaches to the Intentional Perspective on text understanding. It is entitled "Communication and Attitude Revision" (Enclosure 4). This paper, together with Enclosure 3, indicates the vigor with which the discussion on these issues is being conducted at SRI.

1.5 Machine Translation

Also during the period of the previous DARPA contract, Jerry Hobbs and Megumi Kameyama developed an abductive approach to machine translation. Hobbs gave a talk on this, entitled "Machine Translation Using Abductive Inference", at the DARPA Speech and Natural Language Workshop, Asilomar, California, February 1991 (Enclosure 5).

1.6 Efficiency in Abductive Inference

The problem of efficiency in abductive inference systems was addressed by Mark Stickel at a theoretical level in the paper entitled "Upside-Down Meta-Interpretation of the Model Elimination Theorem-Proving Procedure for Deduction and Abduction" (Enclosure 6).

The abductive reasoning procedure in TACITUS is formulated as a top-down reasoning method that reasons from the logical form to be explained backward to facts or assumptions. Such backward reasoning from goals to facts and assumptions provides goal-directedness, which is crucial when using knowledge bases with much information that is irrelevant to explaining any particular logical form. Unfortunately, typical top-down reasoning procedures, such as Prolog and PTTP, have highly redundant search spaces. When explaining a conjunction $A \wedge B$, explanations are sought for B for each explanation of A that is found. Thus, for example, when A and B have no variables in common so that their sets of solutions are independent, B is solved repeatedly with much wasted effort. This is even a more severe problem for abduction than deduction, since allowing assumptions may increase the number of solutions substantially. The problem can be mitigated by adding lemmas or caching, but it is not a trivial task to add such features with adequate performance to a top-down reasoning system.

Bottom-up reasoning systems, such as hyperresolution, reason from facts or assumptions to derived facts. They tend not to be goal directed and can instead be characterized as procedures for generating the inferential closure of a set of facts or assumptions. They often have well developed and effective methods, such as subsumption, to eliminate redundancy. Upside-down meta-interpretation is the process of formulating a top-down reasoning procedure (such as Prolog's input resolution procedure or PTTP's model elimination or linear resolution procedure) for execution by a bottom-up reasoning system, such as hyperresolution with subsumption. The desired goal-orientedness is retained, while the redundancy control methods of the

bottom-up procedure eliminate much of the redundant behavior of the pure top-down reasoning system.

The formula $A \wedge B \rightarrow C$ suggests separate inference rules for top-down and bottom-up execution: from the goal C derive the goals A and B, and from the facts A and B derive the fact C. Upside-down meta-interpretation uses the rules: from the goal C derive the goals A and B, and from facts A and B and the goal C derive fact C. That is, derivation of new facts is contingent on the existence of a matching goal.

This work was presented at a US-Japan Workshop on Theorem Proving at Argonne National Laboratory in June 1991, will be presented at a workshop on Theorem Proving with Analytic Tableaux in Marseilles in March 1993, and has been discussed informally with other researchers on numerous occasions.

In addition, we have implemented a pared-down MiniTACITUS system for use in exploring issues of search efficiency.

1.7 The Semantics of Weighted Abduction

One of the problems with the current abductive framework is the ad hoc character of the weights in the weighted abduction scheme that is used. During the summer of 1992, Clifford Kahn, a graduate student in computer science at Stanford University, worked with Jerry Hobbs on the problem of supplying the weights with a principled semantics. This work was based on Charniak and Shimony's demonstration that a restricted form of weighted abduction for propositional logic can be viewed as the evaluation of a Bayesian network. Kahn's work involved extending this approach to handle first-order predicate logic as well. The work led us to a deeper understanding of the meaning of the weights, although a completely satisfactory solution remains to be worked out.

1.8 Presentations

The "Interpretation as Abduction" framework or selected parts of it were described by Jerry Hobbs in invited talks at the following places:

- The Korea-US Bilateral Workshop on Computers, Artificial Intelligence and Cognitive Science, Seoul, Korea, August 1991.

- Workshop on Natural Language Generation, Borjomi, Republic of Georgia, September 1991.

- Joint Japanese-Australian Workshop on Natural Language Processing, Iizuka, Japan, October 1991.

- Princeton University Cognitive Science Colloquium, November 1991.

- National Centre for Software Technology, Bombay, India, December 1991.

- Center for the Study of Language and Information, Stanford University, Stanford, California, Febr · ry 1992.

- PLUS (Pragmatic Language Understanding System) Workshop, Alghero, Italy, September 1992.

In addition, it formed the basis of the following courses taught by Jerry Hobbs:

- "Abductive Methods of Discourse Interpretation", University of California at Santa Cruz, July 1991.

- "Advanced Computational Linguistics", National Centre for Software Technology, Bombay, India (sponsored by United Nations Development Programme), December 1991.

- "Abductive Methods in Discourse Interpretation", Stanford University, winter quarter, 1992.

Other invited talks about theoretical aspects of research on this project are as follows:

- Jerry R. Hobbs, "Metaphor and Abduction", NATO Workshop on Computational Models of Communication, Trento, Italy, November 1990.

- Douglas Appelt, "Communication and Attitude Revision", NATO Workshop on Computational Models of Communication, Trento, Italy, November 1990.

- Douglas Appelt, "Weighted Abduction for Plan Ascription", Seminar on User Modeling, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Saarbruecken, Germany, November 1990.

- Douglas Appelt, "Communication and Attitude Revision", DFKI Colloquium, Saarbruecken, Germany, November 1990.

- Jerry R. Hobbs, "Machine Translation Using Abductive Inference", DARPA Speech and Natural Language Workshop, Asilomar, California, February 1991.

- Mark Stickel, "Upside-Down Meta-Interpretation of the Model Elimination Theorem-Proving Procedure for Deduction and Abduction", US-Japan Workshop on Theorem Proving, Argonne National Laboratory, June 1991.

- Jerry R. Hobbs, "Metaphor and Abduction", IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts, Implicature, Sydney, Australia, August 1991.

- Mark Stickel, "Upside-Down Meta-Interpretation of the Model Elimination Theorem-Proving Procedure for Deduction and Abduction", to be presented at the Workshop on Theorem Proving with Analytic Tableaux, Marseilles, France, March 1993.

The view of text understanding that we have elaborated is gaining ground world-wide. Among the groups who have adopted this framework in their research are the following:

- The PLUS (Pragmatic Language Understanding System) project in England, France, Netherlands, and Germany.

- A group at Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Saarbrücken, Germany.

- A project on knowledge acquisition at EUROTRA-DK at the University of Copenhagen, Denmark, and the University of Edinburgh in Scotland.

2 Applied Research in Text Understanding

2.1 The TACITUS System

The major impetus to research progress in the period from September 1990 to June 1991 was the effort to prepare the TACITUS system for the evaluation of the system in MUC-3 in May 1991. The principal result of this work was a constellation of techniques for making a full text-understanding system more robust when it encounters all the vagaries of real-world text. This work is described in Enclosure 7.

This research can be divided into three categories:

- Preprocessing
- Syntactic analysis and compositional semantics
- Pragmatics processing, or interpretation

Our progress in each of these areas is discussed in turn.

2.2 Preprocessing

A small amount of preprocessing had been implemented for the TACITUS system prior to 1991, but in 1991, largely because of the demands of the evaluation, a significant amount of work was done in this area. Four principal capabilities were added to the system:

- Spelling Correction. An algorithm for spelling correction was incorporated into the system.
- Hispanic Name Recognition. Because of the frequency of unknown Hispanic names in our data, a statistical trigram model for distinguishing between Hispanic surnames and English words was developed and is used to assign the category `Last-Name` to some of the words that are not spell-corrected.

- Morphological Category Assignment. Words that are not spell-corrected or classified as last names are assigned a category on the basis of morphology. Words ending in "-ing" or "-ed" are classified as verbs. Words ending in "-ly" are classified as adverbs. All other unknown words are taken to be nouns. This misses adjectives entirely, but this is generally harmless, because the adjectives incorrectly classified as nouns will still parse as prenominal nouns in compound nominals. The grammar will recognize an unknown noun as a name in the proper environment.

- A Statistical Relevance Filter was developed by analyzing our training data. We went through the 1300-text MUC-3 development set and identified the relevant sentences. We then determined which unigrams, bigrams, and trigrams were especially diagnostic of relevant sentences. These were used for determining a relevance score for each sentence. A contextually dependent threshold was defined and used to determine whether or not a sentence ought to be processed by the rest of the system. This module gave very good results. We had to process only 20% of the sentences in the test corpus, 56% of the sentences we processed were relevant, and we missed only 18% of the relevant sentences.

We also developed an Anti-Filter based on keywords to capture some of the 18% rejected, but this did not prove successful.

2.3 Syntax and Compositional Semantics

Syntactic analysis and compositional semantics in the TACITUS project are done by the DIALOGIC system. DIALOGIC has perhaps as extensive a coverage of English syntax as any system in existence, including sentence fragments. It produces a logical form in first-order predicate calculus. It includes a well-documented, menu-based component for rapid vocabulary acquisition and a component that produces neutral representations instead of multiple readings for the most common types of syntactic ambiguities, including prepositional phrase attachment ambiguities and very compound noun ambiguities. The DIALOGIC system gives us a large, efficient, reliable, easy-to-use syntactic front-end for any natural language application.

The problem we faced in the MUC-3 effort was that the sentences the system had to deal with were very long, averaging 27 words and sometimes reaching 80 words or more, and exhibited a great deal of syntactic complexity. This led to three principal efforts:

- An agenda-based, scheduling parser was developed making it possible to parse sentences of 60 words or more.

- A recovery method was devised that allowed us to extract most of the propositional content of sentences that failed to parse. The best sequence of grammatical fragments is found, translated into logical form, and passed on to the pragmatics component. The result was that 88% of the propositional content of these sentences was recovered.

- An algorithm we have called "terminal substring parsing" was devised for dealing with very long sentences. The sentence is segmented into substrings, by breaking it at commas, conjunctions, relative pronouns, and certain instances of the word "that". The substrings are then parsed, starting with the last one and working back. For each substring, we try either to parse the substring itself as one of several categories or to parse the entire set of substrings parsed so far as one of those categories. The best such structure is selected, and, for subsequent processing, that is the only analysis of that portion of the sentence allowed. The effect of this technique is to give only short "sentences" to the parser, thereby avoiding the combinatorial explosion, without losing the possibility of getting a single parse for the entire long sentence. This technique was used for sentences longer than 60 morphemes.

A certain amount of work was done on grammar development. About 35 rules were added to the grammar, bringing its size up to 160 rules. Some of these were for handling application-specific syntactic constructions, but most of them were rules that would be useful for written discourse in general, such as some previously unencoded conjunction constructions and propositional attitude adverbials like "he said". A significant amount of debugging was done on the grammar and on the heuristics for choosing among parses of ambiguous sentences. The overall result was that 56% of all sentences parsed with three or fewer errors, and of sentences of under 30 morphemes, 75% parsed with three or fewer errors. Although such statistics are hard to come by, we do not know of any parsing system that has ever performed as well on complex, real-world text. This judgment was validated in the informal grammar evaluation held at the University of Pennsylvania in September 1992, in which DIALOGIC achieved the highest recall score of any evaluated system.

The lexicon was more than doubled in size in preparation for MUC-3,

to about 12,000 words, including about 2000 personal names and about 2000 other proper nouns. Another 8000 words were added before MUC-4, bringing its size to 20,000 words. This expands out to 43,000 morphological variants.

2.4 Pragmatics, or Interpretation

Pragmatics in the TACITUS system is done in accordance with the "Interpretation as Abduction" framework. This requires that the logical form of the sentences in the text be proved, together with the constraints predicates impose on their arguments, allowing for coercions, merging redundancies where possible, and making assumptions where necessary. The TACITUS system employs Mark Stickel's theorem-proving program PTTP (Prolog Technology Theorem Prover), modified to make assumptions and attach costs to proofs, as required by the weighted abduction scheme. The logical form that is produced by the DIALOGIC system is handed over to PTTP to prove abductively in the most economic fashion possible. The system is used for solving the reference, metonymy, syntactic ambiguity, lexical ambiguity, vague predicate resolution, and schema recognition problems occurring in the text.

This method draws inferences from a knowledge base of predicate calculus axioms. For the MUC-3 effort we built up a knowledge base of over 500 axioms, together with a sort hierarchy with over 500 nodes. The pragmatics component was applied to the development sets, and later, of course, to the test sets, and problems that arose were examined intensively.

Abduction can be computationally explosive. It has therefore been a principal concern of ours to devise ways of keeping the reasoning process efficient. We developed three principal techniques: very tight typing of objects in the domain; ordering the search for proofs in such a way that the easiest propositions are generally proved first and the instantiations they produce are propagated to the rest of the goal expression to narrow down the search for its proof; and a strict discipline in how the axioms are written, in particular, avoiding axioms that result in recursion.

In a system that attempts to exploit the redundancy of natural language text, it is very important to recognize implicit identities among entities and at the same time not "over-recognize" identities. We devised several new methods of constraining this aspect of the processing.

2.5 The Results of the MUC-3 Evaluation

The TACITUS system achieved a recall score of 25% on the evaluation, a precision score of 48%. In precision, TACITUS was the best of the 15 systems being evaluated. In recall, it was somewhere in the middle. Our estimate before the evaluation was that we had entered about 25% of the required knowledge into the knowledge base, and our analysis afterwards suggested this was one of the principal factors in the recall score.

2.6 Presentations

We gave the following talks on this aspect of our work:

- Douglas E. Appelt, "The Processing of Naturally-Occurring Texts", DFKI Linguistics Colloquium, Saarbruecken, Germany, November 1990.
- Jerry R. Hobbs, "TACITUS MUC-3 Effort: Analysis of Preliminary Results", Preliminary Meeting, Third Message Understanding Conference, Mountain View, California, February 1991.
- Jerry R. Hobbs, "Site Report: TACITUS", Third Message Understanding Conference, San Diego, California, May 1991.
- Jerry R. Hobbs, "System Summary: TACITUS", Third Message Understanding Conference, San Diego, California, May 1991.
- Jerry R. Hobbs, "Robust Processing of Real-World Natural-Language Texts", Third Conference on Applied Natural Language Processing, Trento, Italy, April 1992.

The material in this phase of our research formed the core of a tutorial that Jerry Hobbs gave with Lisa Rau on three different occasions:

- International Joint Conference on Artificial Intelligence, Sydney, Australia, August 1991.
- AAAI-92 Conference, San Jose, California, July 1992.
- Conference on Information and Knowledge Management, Baltimore, Maryland, November 1992.

3 Applied Research in Information Extraction

3.1 The FASTUS System

The major impetus to progress in system development in 1992 was the effort to prepare for the Fourth Message Understanding Conference (MUC-4) in June 1992. The task of this evaluation was the same as for MUC-3.

We devised a radically new method for information extraction from free text, and implemented it in a system called FASTUS. FASTUS is a (slightly permuted) acronym for Finite State Automaton Text Understanding System. It is a system for extracting information from free text in English, and potentially other languages as well, for entry into a database, and potentially for other applications. It works essentially as a cascaded, nondeterministic finite state automaton. The FASTUS system is described in detail in Enclosure 8.

FASTUS was originally conceived, in December 1991, as a preprocessor for the text-processing system TACITUS, that could also be run in a standalone mode. Most of the design work for the FASTUS system took place during January. The ideas were tested out on finding incident locations and proper names in February. With some initial favorable results in hand, we proceeded with the implementation of the system in March. The implementation of the module for recognizing phrases was completed in March, and the general outline of the module for recognizing patterns was completed by the end of April. On May 6, we did the first test of the FASTUS system on the TST2 set of 100 messages, which had been withheld as a fair test, and we obtained a score of 8% recall and 42% precision. At that point we began a fairly intensive effort to hill-climb on all 1300 development texts in the MUC corpus, doing periodic runs on the fair test to monitor our progress. This effort culminated in a score of 44% recall and 57% precision in the wee hours of June 1, when we decided to run the official test. By the middle of May 1992, it had become clear that the performance of FASTUS on the MUC-4 task was so good that we could make FASTUS not just a pre-processor, but our complete system.

In the actual evaluation, on TST3, we achieved a recall of 44% with precision of 55%, for an F-score of 48.9. On TST4, the test on incidents from a different time span, we observed, surprisingly, an identical recall score of 44%; however, our precision fell to 52%, for an F-score of 47.7. It was reassuring to see that there was very little degradation in performance moving to a time period over which the system had not been trained. These

results were excellent. Out of sixteen systems evaluated, only one system significantly outperformed FASTUS.

More importantly, FASTUS was an order of magnitude or more faster than comparable systems. Other systems of comparable effectiveness required around an hour and a half to process 100 messages. With FASTUS, the entire TST3 set of 100 messages required 11.8 minutes of CPU time on a Sun SPARC-2 processor. The elapsed real time was 15.9 minutes, but observed time depends on the particular hardware configuration involved. To put this into more concrete terms, FASTUS can read 2375 words per minute. It can analyze one text in an average of 9.6 seconds. This translates into 9000 texts per day.

This breakthrough in processing speed results in a corresponding breakthrough in development time, and the combination brings natural language processing to the point of commercial viability.

The operation of FASTUS is comprised of four steps:

1. Triggering
2. Recognizing Phrases
3. Recognizing Patterns
4. Merging Incidents

3.2 Triggering

In the first pass over a sentence, trigger words are searched for. There is at least one trigger word for each pattern of interest that has been defined. Generally, these are the least frequent words required by the pattern. For example, in the pattern

take <HumanTarget> hostage

"hostage" rather than "take" is the trigger word. There are at present 253 trigger words.

In addition, full names are searched for, so that subsequent references to surnames can be linked to the corresponding full names. Thus, if one sentence refers to "Ricardo Alfonso Castellar" but does not mention his kidnapping, while the next sentence mentions the kidnapping but only uses his surname, we can enter Castellar's full name into the template.

3.3 Recognizing Phrases

We will not have systems that reliably parse English sentences correctly until we have encoded much of the real-world knowledge that people bring to bear in their language comprehension. For example, noun phrases cannot be reliably identified because of the prepositional phrase attachment problem. However, certain syntactic constructs can be reliably identified. One of these is the noun group, that is, the head noun of a noun phrase together with its determiners and other left modifiers. Another is what we are calling the "verb group", that is, the verb together with its auxiliaries and any intervening adverbs. Moreover, an analysis that identifies these elements gives us exactly the units we most need for recognizing patterns of interest.

Pass Two in FASTUS identifies noun groups, verb groups, and several critical word classes, including prepositions, conjunctions, relative pronouns, and the words "ago" and "that". Phrases that are subsumed by larger phrases are discarded.

Noun groups are recognized by a 37-state nondeterministic finite state automaton. This encompasses most of the complexity that can occur in English noun groups, including numbers, numerical modifiers like "approximately", other quantifiers and determiners, participles in adjectival position, comparative and superlative adjectives, conjoined adjectives, and arbitrary orderings and conjunctions of prenominal nouns and noun-like adjectives.

Verb groups are recognized by an 18-state nondeterministic finite state machine. They are tagged as Active, Passive, Gerund, and Infinitive. Verbs are sometimes locally ambiguous between active and passive senses, as the verb "kidnapped" in the two sentences

> Several men kidnapped the mayor today.
> Several men kidnapped yesterday were released today.

These are tagged as Active/Passive, and Pass Three resolves the ambiguity if necessary.

Certain relevant predicate adjectives, such as "dead" and "responsible", are recognized, as are certain adverbs, such as "apparently" in "apparently by". However, most adverbs and predicate adjectives and many other classes of words are ignored altogether. Unknown words are ignored unless they occur in a context that could indicate they are surnames.

Lexical information is read at compile time, and a hash table associating words with their transitions in the finite-state machines is constructed. There is a hash table entry for every morphological variant of a word. The

TACITUS lexicon of 20,000 words is used for lexical information. Morphological expansion of these words results in 43,000 morphological variants in the hash table. During the actual running of the system on the texts, only the state transitions accessed through the hash table are seen.

Tests indicate that this module works with better than 95% accuracy.

3.4 Recognizing Patterns

The input to Pass Three of FASTUS is a list of phrases in the order in which they occur. Anything that is not included in a phrase in the second pass is ignored in the third pass. The state transitions are driven off the head words in the phrases. That is, a set of state transitions is associated with each relevant head word-phrase type pair, such as "mayor-NounGroup", "kidnapped-PassiveVerbGroup", "killing-NounGroup", and "killing-GerundVerbGroup". In addition, some nonhead words can trigger state transitions. For example, "bomb blast" is recognized as a bombing.

We implemented 95 patterns for the MUC-4 application. Among the patterns are the following:

killing of <HumanTarget>
<GovtOfficial> accused <PerpOrg>
bomb was placed by <Perp> on <PhysicalTarget>
<Perp> attacked <HumanTarget>'s <PhysicalTarget> with <Device>
<HumanTarget> was injured
<HumanTarget>'s body

As patterns are recognized, incident structures are built up. For example, the sentence

Guerrillas attacked Merino's home in San Salvador 5 days ago
with explosives.

matches the pattern

<Perp> attacked <HumanTarget>'s <PhysicalTarget> in <Location>
<Date> with <Device>

This causes the following incident to be constructed.

Incident:	ATTACK/BOMBING
Date:	14 Apr 89
Location:	El Salvador: San Salvador
Instr:	"explosives"
Perp:	"guerrillas"
PTarg:	"Merino's home"
HTarg:	"Merino"

The incident type is an attack or a bombing, depending on the Device.

3.5 Merging Incidents

As incidents are found, they are merged with other incidents found in the same sentence. Those remaining at the end of the processing of the sentence are then merged, if possible, with the incidents found in previous sentences.

For example, in the sentence

> Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

the phrase

> the terrorist killing of Attorney General Roberto Garcia Alvarado

causes the following incident structure to be built:

Incident:	KILLING
Perp:	"terrorist"
Confid:	–
HTarg:	"Roberto Garcia Alvarado"

while the incident

Incident:	INCIDENT
Perp:	FMLN
Confid:	Suspected or Accused by Authorities
HTarg:	–

is generated from the clause

> Salvadoran President-elect Alfredo Cristiani ... accused the Farabundo Marti National Liberation Front (FMLN)

These two incidents are merged, by merging the KILLING and the INCIDENT into a KILLING, merging "terrorist" and "FMLN" into the more specific "FMLN", and by taking the union of the other slots. The result is the following incident structure.

Incident:	KILLING
Perp:	FMLN
Confid:	Suspected or Accused by Authorities
HTarg:	"Roberto Garcia Alvarado"

Merging is blocked if the incidents have incompatible types, such as a KIDNAPPING and a BOMBING. It is also blocked if they have incompatible dates or locations.

3.6 Future Directions for FASTUS

We are now developing a language that will allow a novice user to be able to begin to specify patterns in a new domain within hours of being introduced to the system. The pattern specification language will allow the user to define structures, to specify patterns either graphically or in regular expressions augmented by assignments to fields of the structures, and to define a sort hierarchy to control the merging of structures.

We would also like to apply the system to a new domain. Our experience with the MUC-4 task leads us to believe we could achieve reasonable performance on a new domain within two months.

Finally, it would be interesting to convert FASTUS to a new language. There is not much linguistic knowledge built into the system. What there is probably amounted to no more than two weeks' coding. For this reason, we believe it would require no more than one or two months to convert the system to another language. This is true even for a language as seemingly dissimilar to English as Japanese. In fact, our approach to recognizing phrases was inspired in part by the bunsetsu analysis of Japanese. We are at present implementing a Japanese version of the phrase recognizer.

FASTUS was more successful than we ever dreamed when the idea was originally conceived. We attribute its success to the fact that its processing

is extremely well suited to the demands of the task. The advantages of the FASTUS system are as follows:

- It is conceptually simple. It is a set of cascaded finite-state automata.
- The basic system is relatively small, although the dictionary and other lists are potentially very large.
- It is effective. Only General Electric's system performed significantly better than FASTUS, and it has been under development for a number of years.
- It has very fast run time. The average time for analyzing one message is less than 10 seconds. This is nearly an order of magnitude faster than comparable systems.
- In part because of the fast run time, it has a very fast development time. This is also true because the system provides a very direct link between the texts being analyzed and the data being extracted.

FASTUS is not a text understanding system. It is an information extraction system. But for information extraction tasks, it is perhaps the most convenient and most effective system that has yet been developed.

3.7 Presentations

The FASTUS system has been described in the following talks:

- Jerry R. Hobbs, "Progress in Text Understanding", TIPSTER Workshop, Philadelphia, Pennsylvania, February 1992.
- Douglas Appelt, "SRI International FASTUS System: MUC-4 Test Results and Analysis", Fourth Message Understanding Conference, Tyson's Corner, Virginia, June 1992.
- Jerry R. Hobbs, "SRI International: Description of the FASTUS System as Used for MUC-4", Fourth Message Understanding Conference, Tyson's Corner, Virginia, June 1992.
- Douglas Appelt, "FASTUS: A System for Extracting Information from Natural Language Text", TIPSTER Workshop, San Diego, California, September 1992.

- Jerry R. Hobbs, "FASTUS: A System for Extracting Information from Natural Language Text", Conference on Information and Knowledge Management, Baltimore, Maryland, November 1992.

We gave a demonstration of the system at the MUC-4 conference, and have given numerous demonstrations of it at SRI. A demonstration version of the system is scheduled to be installed at DARPA headquarters.

4 Other Activities

SRI has been engaging in other activities relevant to text understanding. Jerry Hobbs attended the AAAI Fall Symposium on Knowledge and Action at Social and Organizational Levels at Asilomar, California, November 1991, and delivered a paper entitled "Cognition and Social Action". At the Fifth DARPA Workshop on Speech and Natural Language, Harriman, New York, February 1992, Jerry Hobbs chaired the session on machine translation, and gave a talk on "A National Resource Grammar", which we have been urging as a project for the Linguistic Data Consortium.

5 Future Plans

There are a number of directions in which the research described above needs to be extended, and will be if resources become available. In the theoretical research on the "Interpretation as Abduction" framework three principal efforts are the most urgent:

- Construction of a core knowledge base encoding knowledge about concepts that are relevant to virtually all domains, including the basic facts about space, time, money, and the structure and function of artifacts and organizations. A substantial start was made in this area during our previous DARPA contract, but had to be abandoned due to pressures of the evaluations.

- Implementation and experimentation with the abductive inference process using this knowledge base for interpreting extended discourse with nonliteral and indirect uses of language. Of particular interest in this work would be the discovery of appropriate constraints on interpretations and appropriate heuristics for speeding up the search processes.

- Further work determining the proper semantics for the numbers in the weighted abduction scheme.

We would not, in the immediate future, pursue the development of the TACITUS system. It will be a resource for further development of the FASTUS system, as described below. When our theoretical work is sufficiently mature, we may reconstruct the TACITUS system incorporating our new understandings. But the system that is needed now is a good research vehicle, and the current TACITUS system is too cumbersome to play that role. In addition, if the construction of a National Resource Grammar is funded, we will use it in place of the DIALOGIC component in any text understanding system we build.

Our aim in applied research in information extraction is to incorporate more of the capabilities of the TACITUS system into the FASTUS system. We were surprised how much we could do with the finite-state technology. We have come nowhere near the limits of the technology. We perceive certain analogies between the TACITUS-style and FASTUS-style systems.

- The phrase recognition component of FASTUS implements the reliable portion of the syntactic analysis of TACITUS.
- Many of the axioms encoding the knowledge in TACITUS can be recast as patterns to be recognized in Phase 3 of FASTUS's operation.
- The factoring performed by TACITUS to resolve co-reference is very much like the step in FASTUS of merging incidents.

We believe these similarities can be exploited for moving capabilities from the TACITUS system to the very much faster FASTUS system.

Our plans for the MUC-5 evaluation, to be held in July 1993, are to spend the month of March rapidly building up FASTUS's performance to the level of the leading TIPSTER systems. We believe we can do this, because we were able to do essentially that in May 1992 preparing for MUC-4, and now we have more convenient interface tools. The remaining months will be spent in a concerted effort to break through that glass ceiling that seems to limit message processing systems to at most 60% recall and 60% precision.

References

[1] Appelt, Douglas E., 1991. "Discourse Processing in the TACITUS System", *Proceedings*, Third Message Understanding Conference, San Diego, California, May 1991.

[2] Appelt, Douglas E., 1992. "Communication and Attitude Revision", in A. Ortony, J. Slack, and O. Stock, eds., *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, Springer, Heidelberg. (Enclosure 4)

[3] Appelt, Douglas, John Bear, Jerry R. Hobbs, David Israel, and Mabry Tyson, "SRI International FASTUS System: MUC-4 Test Results and Analysis", *Proceedings*, Fourth Message Understanding Conference, Tyson's Corner, Virginia, June 1992, pp. 143-147.

[4] Hobbs, Jerry R., 1991. "Machine Translation Using Abductive Inference", *Proceedings*, DARPA Speech and Natural Language Workshop, Asilomar, California, February 1991. (Enclosure 5)

[5] Hobbs, Jerry R., 1991. "Site Report: TACITUS", *Proceedings*, Third Message Understanding Conference, San Diego, California, May 1991.

[6] Hobbs, Jerry R., 1991. "System Summary: TACITUS", *Proceedings*, Third Message Understanding Conference, San Diego, California, May 1991.

[7] Hobbs, Jerry R., 1991. "Interpretation as Abduction", Proceedings of the Korea-U Bilateral Workshop on Computers, Artificial Intelligence, and Cognitive Science, Seoul, Korea, August 1991, pp. 207-234.

[8] Hobbs, Jerry R., 1992. "Metaphor and Abduction", *Proceedings*, IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts, Implicature, Sydney, Australia, August 1991, pp. 52-61.

[9] Hobbs, Jerry R., 1992. "Metaphor and Abduction", in A. Ortony, J. Slack, and O. Stock, eds., *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, Springer, Heidelberg, pp. 35-58. Also published as SRI Technical Note 508, SRI International, Menlo Park, California. August 1991. (Enclosure 2)

[10] Hobbs, Jerry R., 1992. "A Unified Abductive Treatment of the Intentional and Informational Aspects of Discourse Interpretation: A Preliminary Report", manuscript. (Enclosure 3)

[11] Hobbs, Jerry R., Douglas Appelt, Mabry Tyson, John Bear, and David Israel, "SRI International: Description of the FASTUS System as Used for MUC-4", *Proceedings*, Fourth Message Understanding Conference, Tyson's Corner, Virginia, June 1992, pp. 268-275.

[12] Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, and Mabry Tyson, 1992. "FASTUS: A System for Extracting Information from Natural-Language Text", SRI Technical Note 519, SRI International, Menlo Park, California, November 1992. (Enclosure 8)

[13] Hobbs, Jerry R., Douglas E. Appelt, John Bear, Mabry Tyson, and David Magerman, "The TACITUS System: The MUC-3 Experience", Technical Note 509, Artificial Intelligence Center, SRI International, Menlo Park, California, September 1991.

[14] Hobbs, Jerry R., Douglas E. Appelt, John Bear, Mabry Tyson, and David Magerman, 1992. "Robust Processing of Real-World Natural-Language Texts", *Proceedings*, Third Conference on Applied Natural Language Processing, Trento, Italy, April 1992, pp. 186-192.

[15] Hobbs, Jerry R., Douglas E. Appelt, John Bear, Mabry Tyson, and David Magerman, 1992. "Robust Processing of Real-World Natural-Language Texts", in *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, P. Jacobs, editor, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 13-33. (Enclosure 7)

[16] Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. "Interpretation as Abduction", to appear in *Artificial Intelligence Journal*. (Enclosure 1)

[17] Stickel, Mark, 1993. "Upside-Down Meta-Interpretation of the Model Elimination Theorem-Proving Procedure for Deduction and Abduction", to appear in the *Journal of Automated Reasoning*. Also published as a Technical Report, Institute for New Generation Computer Technology, Tokyo, Japan, May 1991. (Enclosure 6)